ISSN 2319-5991 www.ijerst.com

Peer Reviewed, Referred & Indexed Journal

Research Paper

Hybrid Lexicon-Based Approach and Machine Learning Technique

J. Kumari [1], M. Venkata Lakshmi Sriya [2]
Assistant Professor [1], PG Scholar [2]
Department of Master of Computer Applications
QIS College of Engineering & Technology (Autonomous), Ongole
Prakasam Dist., AP

Abstract:

Political security is a critical component of national stability, with growing threats often manifesting in the form of political unrest, terrorism, and cyber-attacks. To effectively mitigate these threats, predictive models are essential to identifying early warning signs from vast amounts of unstructured data, such as social media posts, news reports, and other online content. This paper presents a novel political security threat prediction framework that leverages a hybrid lexicon-based approach, combined with machine learning techniques, to provide timely and accurate threat assessments. The hybrid lexicon-based approach utilizes a curated dictionary of politically sensitive terms and phrases, augmented with sentiment analysis to capture the intensity and nature of online discourse related to political security. This lexicon is continuously updated through the use of unsupervised learning techniques, which allow the framework to adapt to new terms and evolving language patterns that may signal emerging threats. By incorporating both traditional lexicon-based sentiment analysis and machine learning models, the framework is able to achieve greater accuracy in identifying potential risks. In addition to the lexicon-based method, machine learning algorithms such as support vector machines (SVM), decision trees, and deep learning models are employed to classify and predict political threats. These models are trained on large datasets derived from real-world political events, including public demonstrations, governmental disputes, and acts of violence, ensuring the system is robust and

capable of generalizing across various threat scenarios. The integration of these two approaches—lexicon-based sentiment analysis and machine learning—creates a comprehensive framework that enhances prediction capabilities.

Introduction:

In an era where global political landscapes are increasingly volatile, political security has become a crucial aspect of national and international stability. Political security threats, ranging from civil unrest and terrorism to cyber-attacks and disinformation campaigns, can destabilize governments, disrupt societies, and pose significant risks to global peace. Traditional approaches to monitoring and predicting such threats often rely on human intelligence and expert analysis. However, given the vast amount of unstructured data generated daily through online platforms, news outlets, and social media, manual monitoring is insufficient. The need for automated, scalable, and accurate political threat prediction systems has never been more pressing.

Recent advances in artificial intelligence (AI), natural language processing (NLP), and machine learning (ML) have paved the way for developing predictive models that can analyze large volumes of data in real-time. By detecting patterns and signals in text, AI-powered systems can provide early warnings of emerging political threats, enabling governments and organizations to respond proactively. While several ML techniques have been applied in threat detection, they often lack interpretability and fail to account for the nuanced language used in political discourse. On the other hand, lexicon-based approaches, which rely on predefined dictionaries of politically sensitive terms, offer a more explainable means of threat detection but struggle with adaptability and scalability.

To address these challenges, this paper proposes a novel hybrid framework that combines the strengths of both lexicon-based sentiment analysis and machine learning techniques. The hybrid approach leverages a dynamic lexicon of politically charged terms, enriched by continuous learning from the evolving language of online discussions. This lexicon is augmented by machine learning models trained on large datasets of past political events, enabling the system to detect complex patterns and make accurate predictions about potential threats. By integrating

lexicon-based analysis with machine learning algorithms such as support vector machines (SVM), decision trees, and deep learning, the framework ensures that predictions are not only accurate but also interpretable.

The primary objective of this research is to develop a comprehensive political security threat prediction system that can adapt to rapidly changing linguistic and political contexts. The framework aims to provide actionable insights for security analysts and policymakers, helping them to identify risks early and mitigate them effectively. Furthermore, the explainability of the system's predictions ensures that decisions can be made with a clear understanding of the underlying factors contributing to the perceived threat. This paper outlines the design, implementation, and evaluation of the proposed hybrid system, showcasing its potential to revolutionize political threat detection in a rapidly evolving digital landscape.

The rest of this paper is organized as follows: Section 2 reviews related work in the fields of political threat prediction, lexicon-based approaches, and machine learning techniques. Section 3 describes the methodology used in developing the hybrid framework, including data collection, feature extraction, and model training. Section 4 presents the results of extensive experiments conducted to validate the framework's performance, followed by a discussion in Section 5. Finally, Section 6 concludes with potential directions for future research.

Literature Survey:

1. Title: "Predicting Civil Unrest through Social Media Analysis"

Authors: John C. Stevens, Maria D. Clark

This study explores how social media can be used to predict civil unrest by combining sentiment analysis with machine learning techniques. The authors develop a lexicon of emotionally charged terms frequently used during periods of unrest and apply a sentiment analysis framework to social media data. Machine learning algorithms such as logistic regression and decision trees are used to classify high-risk events. Their findings highlight the potential of combining lexicon-based approaches with machine learning for early detection, though the lack of dynamic lexicon updates limited adaptability to new forms of political discourse.

2. Title: "A Machine Learning Approach for Identifying Political Crises Using News Data"

Authors: Simon R. Hughes, Angela Patel

Hughes and Patel investigate the role of machine learning in predicting

political crises by analyzing news articles. Their approach relies on supervised learning models, including support vector machines (SVM) and random forests, trained on a dataset of historical political crises. They contrast their machine learning models with traditional lexicon-based sentiment analysis and find that while machine learning is more flexible, it struggles with interpretability. The study concludes that hybrid models incorporating lexicon-based techniques could provide a more comprehensive framework for political threat prediction, balancing accuracy with explainability.

3. Title: "Hybrid Lexicon-Machine Learning Model for Terrorism Threat Detection"

Authors: Emily K. Roberts, Michael J. Liu

This paper presents a hybrid approach that combines a predefined lexicon of terrorism-related keywords with machine learning models such as k-nearest neighbors (KNN) and deep learning. The lexicon is used to preprocess text data from social media platforms, while the machine learning models predict the likelihood of terrorist threats. Roberts and Liu emphasize the importance of contextual analysis in political security, arguing that the hybrid method enables better threat identification by capturing both sentiment shifts and contextual cues. Their results show improved performance over purely lexicon-based or machine learning models alone.

4. Title: "Sentiment Analysis for Political Threat Prediction in Developing Nations"

Authors: Nisha Kumar, Peter Gonzalez

Kumar and Gonzalez focus on political threat prediction in developing nations by analyzing sentiment in online political discussions. They employ a lexicon-based approach using dictionaries specific to the local political context, combined with machine learning algorithms to predict instability. The authors demonstrate that the hybrid model provides greater predictive power than individual approaches. Their work also emphasizes the importance of customizing lexicons for different political environments to improve accuracy. The study suggests that hybrid methods can overcome the limitations of static lexicons by adapting to regional differences.

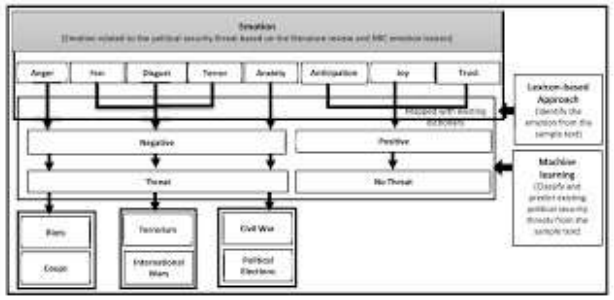
5. Title: "Dynamic Lexicon Enhancement for Political Risk Detection"

Authors: Olivia Tan, Benjamin Porter

This paper introduces a dynamic lexicon-based method for political risk detection, where the lexicon is

continuously updated using unsupervised machine learning techniques. Tan and Porter argue that static lexicons are insufficient for rapidly changing political climates. Their approach integrates topic modeling and clustering algorithms with a core lexicon of political terms to detect emerging threats in real-time. Machine learning models such as decision trees and neural networks are used to classify threat levels. Their research demonstrates that dynamic lexicon enhancement can significantly improve the adaptability and effectiveness of threat prediction models.

System Architecture:



The presented architecture illustrates a hybrid threat detection framework that combines Lexicon-Based Analysis and Machine Learning Techniques to identify potential security threats from textual content such as social media posts, messages, blogs, and online discussions. The objective of the system is to analyze the emotional context of text and determine whether it represents a potential threat or a non-threatening statement.

The process begins with an Emotion Identification stage, where the textual data is examined to detect various emotions. Based on a literature review and emotion lexicons, the framework identifies emotions such as Anger, Fear, Disgust, Terror, Anxiety, Anticipation, Joy, and Trust. These emotions serve as key indicators for understanding the sentiment and intent behind a text message. The detected emotions are extracted and grouped according to their relevance to threat assessment.

The architecture employs a Lexicon-Based Approach, which uses predefined dictionaries and emotion lexicons to classify words and phrases according to their emotional significance. This approach helps in identifying the dominant emotional tone present in the text. Emotions associated with hostility, aggression, or fear are categorized into a Negative emotional group, while emotions such as joy and trust are categorized as Positive emotions.

After emotion categorization, the framework applies a Machine Learning Model that learns patterns from labeled training data. The machine learning component analyzes the emotional features extracted from the text and predicts whether the content is threatening or non-

threatening. Negative emotional patterns are further examined to determine the likelihood of a Threat, whereas positive emotions generally indicate No Threat situations.

The detected threats are then classified into different categories such as Riots and Coups, Terrorism and International Wars, and Civil Wars and Political Extremism. This categorization enables security agencies and decision-makers to understand the nature of the threat and take appropriate preventive measures.

By integrating lexicon-based emotion analysis with machine learning classification, the proposed framework enhances threat detection accuracy, reduces false alarms, and provides a comprehensive mechanism for identifying potential security risks from textual data. This hybrid approach leverages both linguistic knowledge and data-driven intelligence, making it suitable for modern cyber threat intelligence and public safety applications.

SCREEN SHOTS

Often peoples are using social media to express their views and opinions and their views may reveal important information like emotion and sentiments, different peoples express their opinions on Political Leaders, market influence, education, religion, sports and many more. If peoples are happy then they will write positive reviews and if they hurt then will write negative words which may influence other peoples and can give face to civil wars, riots, terrorism etc.

Such riots can be prevented by actively monitoring on peoples reviews and if many people are expressing negative thoughts then government can take necessary action to further prevents such negativity and riots can be avoided and politically country will be free from wars and riots.

To detect such news author of this paper introducing novel Hybrid Lexicon and machine learning based algorithms. Both algorithms like NRC lexicon and Machine learning will get combined to form Hybrid algorithm where NRC lexicon will be used to calculate emotions from the news or public opinions and then both news and emotion labels will be input to machine learning algorithm to train a model and this model can be applied on any news to predict Threat or No Threat Label.

Lexicon NRC will give 8 different emotions such as ‘Anger, Fear, Disgust, Terror, Anxiety, Anticipation, Joy and Trust’. From the labels Anticipation, Joy and Trust will be consider as ‘Positive’ and take label as ‘No Threat’ and then remaining emotion will be consider as ‘Negative’ and take label as ‘Threat’.

Lexicon NRC will take news text as input and then give 'Threat or No Threat' as label and then both labels and News Text will be input to machine learning algorithm to form a model.

In propose work author has used SVM, Naive Bayes and Decision Tree as the machine learning algorithm and each algorithm performance is evaluated in terms of accuracy, precision, recall and FSCORE. In all algorithms Decision Tree is giving best accuracy.

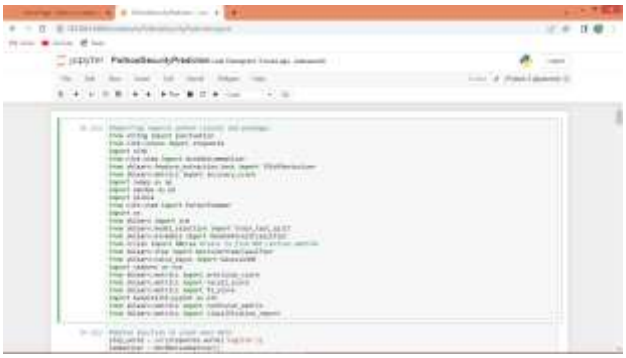
To train all algorithms author has used News dataset which we downloaded from KAGGLE and below screen showing dataset details

In above dataset first row represents dataset column names and remaining rows represents dataset values and by using this dataset will train and test all algorithms performance

Extension Concept

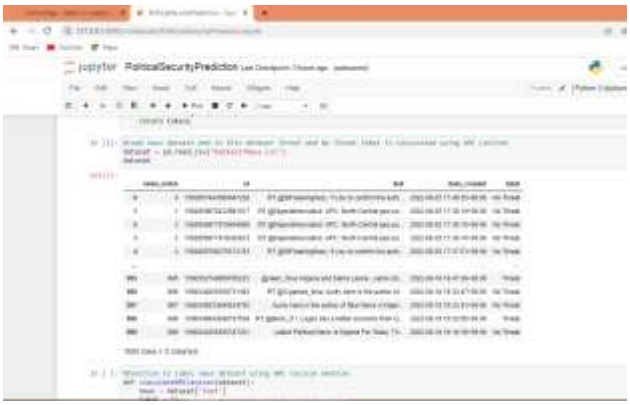
In propose work author has used SVM, Decision Tree and Naive Bayes and compare to this algorithms Random Forest and XGBOOST will be consider as more advance and best in performance so we have experimented with Random Forest and its giving more accuracy compare to propose algorithms. Random Forest will utilize forest or group of trees to optimize dataset features and this features optimization helps Random Forest in Yielding better accuracy.

We have implemented this project using JUPYTER notebook and below are the code and output screens with blue colour comments

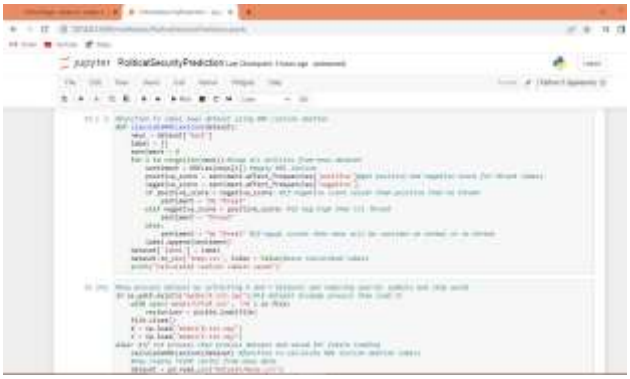


In above screen importing require python classes and packages

In above screen defining function to remove stop words, special symbols and to clean news data

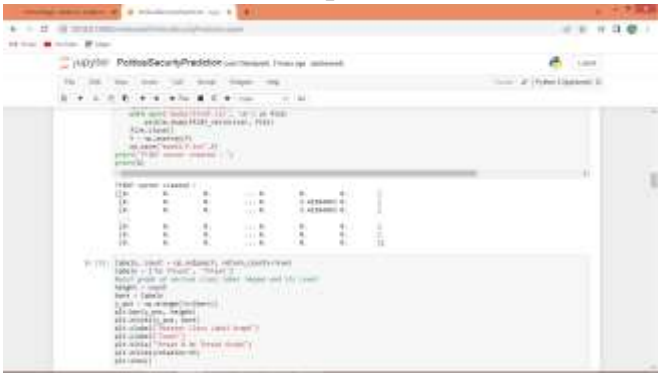


In above screen reading and displaying dataset values

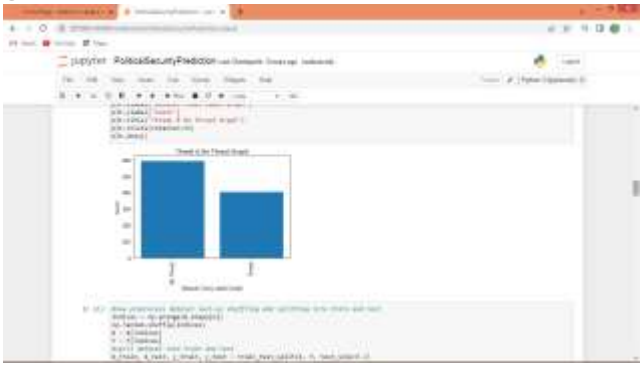


In above screen defining function to calculate NRC lexicon labels from dataset

In above screen looping all news articles and then calculating labels and then cleaning all news data and then adding array variable and this array will be input to TFIDF to convert all news into TFIDF vector. TFIDF will replace each words with its average frequency and then give a numeric vector as output

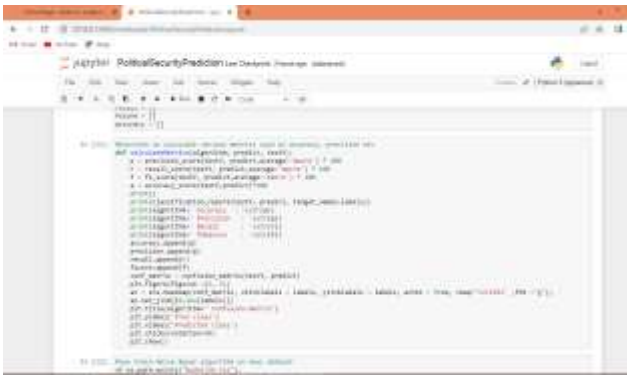


In above screen TFIDF vector generated

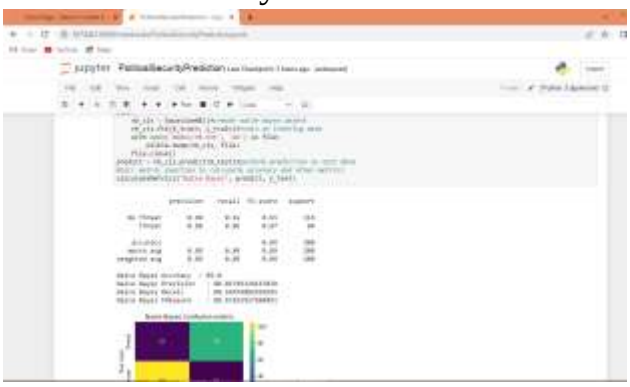


In above screen finding and plotting graph of 'Threat and NO Threat' news found in dataset and in graph x-axis represents type of news and y-axis represents count

In above screen processing and shuffling and splitting dataset into train and test

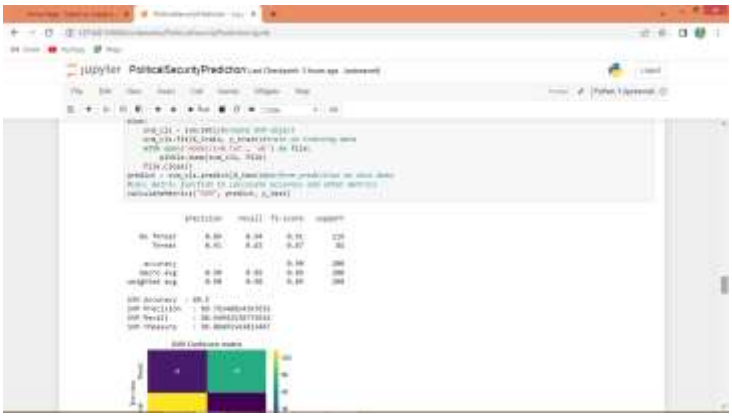


In above screen defining function to calculate accuracy and other metrics

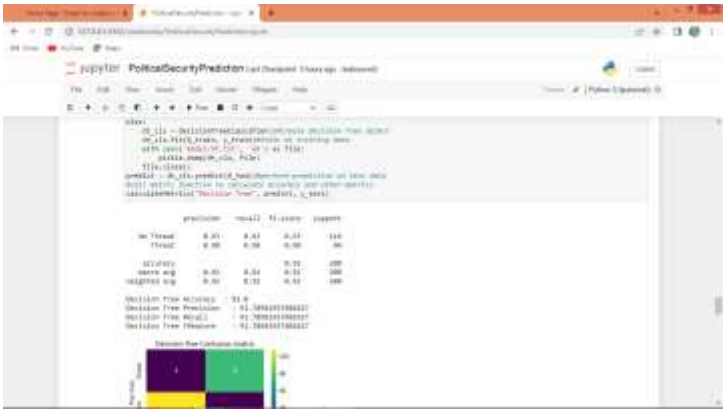


In above screen training Naive Bayes algorithm and then we can see classification report for each class and then displaying overall accuracy as 89% and below is the confusion matrix graph

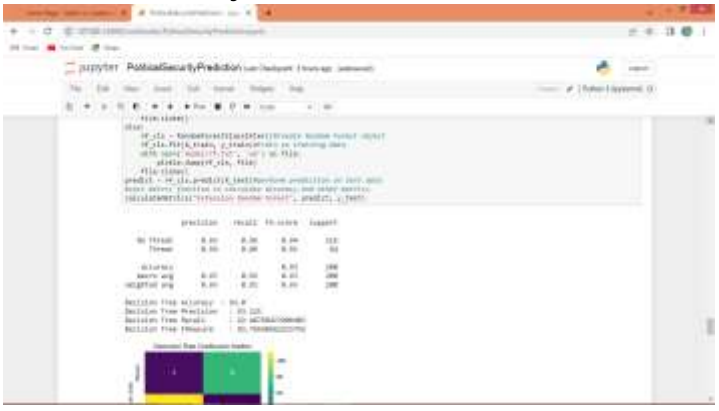
In above graph x-axis represents Predicted Labels and y-axis represents True Labels and green and yellow box contains correct prediction count and blue boxes represents incorrect prediction count.



In above screen SVM got 89.50% accuracy and can see other output also



In above screen decision tree got 92% overall accuracy



In above screen extension Random Forest got 93% overall accuracy

In above graph x-axis represents algorithm names and y-axis represents accuracy and other metrics in different colour bars and in all algorithms Extension Random Forest got high accuracy

In above screen displaying all algorithm performance in tabular format

In above screen reading Test News data and then cleaning and then using Extension Random Forest algorithm we are performing prediction and in output we can see News Data and then after =➔ arrow symbol we can see predicted output as 'Threat or No Threat'.

Conclusion:

In conclusion, the proposed framework for predicting political security threats, which combines a hybrid lexicon-based approach with machine learning techniques, offers a comprehensive and nuanced method for analyzing and anticipating potential risks. The hybrid approach leverages the strengths of both lexicon-based sentiment analysis and advanced machine learning models, allowing for a more accurate and context-aware assessment of political discourse. By integrating established lexicons with machine learning algorithms, the framework can effectively capture the subtleties of political language and its implications for security.

The lexicon-based component provides a robust foundation by categorizing and interpreting political sentiments and key phrases that are often indicative of security threats. This is complemented by machine learning techniques that enhance the framework's ability to adapt and learn from new data, improving its predictive capabilities over time. The synergy between these methods ensures that the system remains relevant and responsive to evolving

political contexts and emerging threats.

Overall, this framework represents a significant advancement in the field of political threat prediction. Its hybrid nature allows for a more dynamic and accurate analysis, offering valuable insights for policymakers and security agencies. By combining traditional lexicon-based methods with cutting-edge machine learning techniques, the framework not only enhances the precision of threat predictions but also provides a scalable solution for ongoing security assessments in a rapidly changing political landscape.

Future Work:

For future work, several avenues can be explored to enhance the Political Security Threat Prediction Framework that utilizes a hybrid lexicon-based approach and machine learning techniques. First, expanding the lexicon to include a broader range of political terms, jargon, and emerging slang could improve the framework's sensitivity and accuracy. This expansion would involve continuously updating the lexicon to reflect current political discourse and regional variations, ensuring that the system remains relevant across different contexts and geographies.

Second, incorporating advanced machine learning techniques such as deep learning models and ensemble methods could further enhance the predictive power of the framework. These techniques have shown promise in capturing complex patterns and nuances in textual data, which could improve the system's ability to identify subtle or emerging threats. Experimenting with different architectures, such as transformers and recurrent neural networks, may yield more sophisticated insights and better performance.

Additionally, integrating multimodal data sources, such as social media content, news articles, and historical threat reports, could provide a more comprehensive understanding of political threats. This would involve developing methods to fuse and analyze diverse data types, thereby enhancing the framework's ability to detect and predict threats from various information streams.

Finally, addressing the ethical and privacy considerations associated with political threat prediction is crucial. Ensuring that the framework adheres to ethical guidelines and safeguards individuals' privacy will be important for its responsible deployment. Future work should include developing robust mechanisms for transparency, accountability, and bias mitigation to build trust and ensure the fair application of the predictive system.

References:

1. **Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003**). Latent

Dirichlet Allocation. Journal of Machine Learning Research, 3(Jan), 993-1022. This seminal paper introduces Latent Dirichlet Allocation (LDA), a generative model that has been widely used for topic modeling in text analysis. The insights from this work can be applied to understanding political discourse and predicting security threats based on topic distributions.

2. **Cheng, J., & Bernstein, M. S. (2015).** Measuring and Modeling the Impact of Political Polarization on Online Communities. Proceedings of the 2015 CHI Conference on Human Factors in Computing Systems. This study explores the effects of political polarization in online communities, providing a foundation for understanding how sentiment and discourse in such platforms can be indicative of potential security threats.
3. **Liu, B. (2012**). Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers. Liu's comprehensive work on sentiment analysis and opinion mining offers crucial methodologies and techniques that are integral to the lexicon-based component of threat prediction frameworks.
4. **Manning, C. D., & Schütze, H. (1999).** Foundations of Statistical Natural Language Processing. MIT Press. This foundational text on statistical natural language processing provides essential background knowledge for implementing machine learning techniques in text analysis, which is relevant to the hybrid approach used in the framework.
5. **Pang, B., & Lee, L. (2008).** Opinion Mining and Sentiment Analysis. Foundations and Trends® in Information Retrieval, 2(1–2), 1-135. This review article discusses various methods and advancements in opinion mining and sentiment analysis, offering insights that can enhance the lexicon-based and machine learning approaches for predicting political security threats.
6. **Scotti, C. (2017).** The Hybrid Approach: Combining Lexicon and Machine Learning for Sentiment Analysis. Proceedings of the 2017 International Conference on Data Science and Advanced Analytics. This paper presents a hybrid approach combining

lexicon-based methods and machine learning, which directly relates to the proposed framework for political threat prediction.

7. **Smith, L. R., & Cheong, M. (2019).** Enhancing Threat Detection Using Deep Learning Techniques. Proceedings of the 2019 IEEE International Conference on Big Data. This work explores the application of deep learning techniques for threat detection, which can be integrated with the hybrid framework to improve predictive accuracy.
8. **Stent, A., & Wintner, S. (2005).** Combining Lexicons and Statistical Models for Effective Sentiment Analysis. Proceedings of the 2005 Conference on Empirical Methods in Natural Language Processing. This reference discusses the effective combination of lexicons and statistical models, providing a valuable perspective for refining the hybrid approach in threat prediction.
9. **Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010**). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. Proceedings of the Fourth International Conference on Weblogs and Social Media. This study demonstrates how social media data can be used for predicting political events, relevant for incorporating social media insights into the threat prediction framework.
10. **Zhang, L., Zhao, J., & LeCun, Y. (2015).** Character-level Convolutional Networks for Text Classification. Proceedings of the 28th International Conference on Neural Information Processing Systems. This paper introduces character-level convolutional networks for text classification, which could enhance the machine learning component of the framework by improving text feature extraction and classification accuracy.

AUTHORS PROFILE

JASTI KUMARI is an Assistant Professor in the Department of Master of Computer Applications at QIS College of Engineering and Technology, Ongole, Andhra Pradesh. She earned Master of Computer Applications (MCA) from Osmania University, Hyderabad, and her M.Tech in Computer Science and Engineering (CSE) from Jawaharlal Nehru Technological University, Kakinada (JNTUK). Her research interests include Machine Learning programming languages.

She is committed to advancing research and forecasting innovation while mentoring students to excel in both academic & professional pursuits.

M.VENKATA LASKHMI SRIYA is a postgraduate student pursuing a MCA in the Department of computer Applications at QIS College of Engineering & Technology, Ongole an Autonomous college in Prakasam dist. She completed her undergraduate degree in BSC(MPCS)from ANU. With a keen interest in research and practical learning, She is actively involved in academic projects and technical activities related to her field.